Identifying coherent subnetworks in genome scale metabolic networks

Verwoerd, W.S.

Centre for Advanced Computational Solutions, AGLS Division, PO Box 84, Lincoln University, Canterbury, New Zealand
Email: verwoerw@lincoln.ac.nz

Keywords: *metabolic network, biochemical pathway, flavonoid, constraint, subnet*

EXTENDED ABSTRACT

Constraint-based methods to analyse metabolic networks require the classification of metabolites into external compounds that are exchanged with the system environment, and internal ones that form intermediate steps on the metabolic pathways and which are therefore subject to stoichiometry constraints in the steady state.

For eukaryotic cells it is not practical to analyse the complete cellular network, and so the subnetwork that describes the specific phenomena under study needs to be extracted. This study focuses on flavonoid production in the model plant *Arabidopsis Thaliana*.

Software was developed to extract the full *Arabidopsis* metabolic network from the AraCyc database, in a format required by standard network analysis packages. Considerable processing is also required to reconcile information about reaction directions from different database tables and to classify metabolites as internal or external to the subnetwork. The paper outlines the strategies implemented in the software to address these issues.

For compiling a subnetwork, the first step was to identify a prototype subnetwork by collecting all empirical reaction pathways that are known to be associated with flavonoids, using a keyword search. A subsequent manual procedure was developed to extend the prototype step by step until all external metabolites of the subnetwork are accounted for as belonging to the set of universal exchange compounds such as nutrients, nucleotides and amino acids.

The resulting subnetwork for flavonoids in Arabidopsis contains 115 metabolites and 89 reactions, which is sufficiently compact for the calculation of elementary modes and subsequent analysis.

A feature of the approach used here is that it is not based just on network topology, as is the case for most standard algorithms based on graph theory concepts. First, the prototype network used as starting point, is identified using empirically compiled metabolic pathways as additional input. Also, an explicit listing of compounds that are acceptable as external compounds based on biochemical knowledge of their general availability is used in addition. Nodes corresponding to these are suitable locations for the network to be cut in order to separate the subnetwork.

Previous work attempted to identify such nodes directly from the network structure, for example based on network connectivity of compounds. In the present context, however, this was not adequate to isolate a coherent subnetwork containing the prototype. Nevertheless it is found that the explicit identification of external compounds combined with the network separation algorithm used previously, does produce a subnetwork compatible with the one found from the method presented here.

1. INTRODUCTION

Constraints based modelling has become a major tool in the systems biology approach to analyse the biochemical network that describes metabolism processes in cells. This is also referred to as flux balance analysis (FBA). A detailed exposition of this approach is found in a recent book by Palsson (2006). Implementations of the method are available in software packages such as Cellnet Analyzer (CNA) by Klamt et al. (2007) and YANA by Schwarz et al. (2005).

The structure of the network is given by specification of the stoichiometry matrix **S**, in which each column represents a chemical reaction, and each row a metabolite compound. The matrix elements are the numerical stoichiometry coefficients which are integer numbers and are negative for reactants (conventionally called substrates) and positive for reaction products.

The dynamics of the changes in the concentrations of metabolites is determined by conservation of atomic species (or mass) expressed by the equation

$$\frac{d\mathbf{x}}{dt} = \mathbf{S} \cdot \mathbf{v} \qquad (1.1)$$

Here **x** is a vector of metabolite concentrations, and **v** is a vector of fluxes through each equation. The flux is basically the number of molecular "copies" of the reaction taking place per unit time in a unit volume.

Assuming that chemical equilibria are established quickly compared to the rate at which the external or regulatory environment of a cell changes, cellular processes can be regarded as transitions between steady biochemical states. Such a steady state is characterised by the flux through each chemical reaction.

For a network of *N* reactions, any point in an *N*-dimensional flux space hence describes a steady network state; but not all points in the space are feasible. Equation (1.1) shows that steady state flux vectors are eigenvectors of **S** with a zero eigenvalue, i.e. they belong to the (left) null space of **S**. Moreover since fluxes cannot be negative, the feasible states have to lie in a convex subspace of the null space. This subspace is described by its set of convex edge vectors, also known as *extreme pathways*, or by closely related *elementary modes*.

A metabolic network can be graphically represented as a reaction map, in which the compounds form the nodes and reactions are represented by the connecting links, i.e. directed edges. Stoichiometry constraints are in principle mass balances, and apply at all internal nodes in the network. However, the boundary of the network is defined by nodes that represent compounds freely exchanged with the environment, such as water and nutrients. These are considered as reservoirs not subject to mass conservation. This partitions the stoichiometry matrix into internal and exchange blocks.

The partitioning clearly applies to the metabolic network as a whole, since a cell exists in an environment where a limited number of relatively simple nutrient compounds are supplied and some waste compounds are delivered into the environment.

However, it may be possible to further subdivide the network into distinct functional sections that only exchange a limited number of compounds between themselves. For example, catabolism and anabolism are conventionally considered as the main metabolic processes, and a set of only 11 compounds are believed to be transmitted between the collections of reactions that make up each of those main blocks [Atkinson (1977)]. So if these 11 compounds are also designated as external, not explicitly subject to mass conservation, either the catabolism or anabolism subnetworks could be isolated and studied on its own e.g. to determine its steady state flux balances.

The subject of this article is to explore how such a subnetwork representing some particular functional aspect of the complete network can be determined, in a way that is consistent with both the network topology and biochemical or biological knowledge. In doing so, it has to be taken into account that present knowledge of the metabolic networks for most organisms is still incomplete and that existing metabolic databases focus on individual pathways and reactions so that the identification of external compounds for the known network also requires analysis.

The problem of determining a subnetwork is reminiscent of standard problems in graph theory, such as cluster determination, for which efficient algorithms such as Markov clustering [Enright et al. (2002)] are available. Since an equation (edge) can impinge on more than two compounds (vertices) a metabolic network is however not a simple graph for which such algorithms are designed, but instead a so-called hypergraph. Moreover, a subnetwork is also not a cluster as the term is usually defined. It is defined not by the close linkage of a group of nodes, but rather by the fact that the group of nodes is connected to the rest

of the network only by members of a specific subset of nodes – the external compounds. An earlier method that relies on network topology (connectivity numbers) to identify this subset, is to be compared in section 3 with the strategy proposed here.

In section 2 the procedure followed in this work to extract the full stoichiometry matrix and its partitioning from a standard metabolic database is outlined. Section 3 describes a strategy to isolate a subnetwork, and applies this to obtain the flavonoid secondary metabolism subnetwork for the model plant *Arabidopsis Thaliana*.

2. EXTRACTING THE STOICHIOMETRY MATRIX

From the genome sequence of an organism, enzyme assignments can be made and hence associated reactions identified by comparison with known enzyme activities in other organisms. This yields a putative genome scale metabolic network, that has to be curated by manual inspection and comparison with published experimental work. This formidable task is still at various stages of completion depending on the organism, but for *Arabidopsis* it has been taken to a fairly advanced stage by TAIR [Rhee et al. (2003)]. The resulting AraCyc database is an implementation of the generic BioCyc database system [Karp et al. (2005)] and is available as a series of flatfile tables describing compounds, reactions, enzymes and pathways. Version 3.0 of AraCyc was used for this study and encompasses about 1500 reactions and 1200 compounds.

In principle extraction of **S** matrix element values from AraCyc is a straightforward parsing problem and has been automated in a software system, consisting of a series of AWK scripts, for this project. However there are a few pitfalls to avoid.

Firstly, a substantial number of reactions in AraCyc are not chemically balanced. Most of these are merely placeholders for reaction classes or generic reactions in which e.g. one reactant might be given as "a fatty acid". In a few cases all reactants may not yet be known. As an unbalanced equation does not constitute a valid stoichiometric constraint, the balance state of all reactions are checked by the software and only balanced reactions retained in the **S** matrix. The resulting omission of some partial knowledge of the network, will hopefully be alleviated as updated versions of AraCyc (currently appearing about twice a year) progressively eliminate unbalanced reactions from the database.

Secondly, information about reaction directions (that determines the arithmetic sign of matrix elements) is distributed in the database and sometimes ambiguous. The explicitly shown direction in the reactions table is merely stated according to biochemical conventions. Under cellular conditions of temperature, concentration levels etc. the reaction may well proceed in the opposite direction or even be reversible. Reversible reactions are specially treated by elementary mode analysis algorithms, so this information needs to be recorded along with the **S** matrix. To complicate matters, the relationship between reactions and enzymes is a many-to-many relationship. In order to reconcile the reaction direction information stored as part of the reaction specification with that stored in the enzyme and pathway tables, the following strategy has been implemented in the parsing software.

If there is only one enzyme for a reaction, and a direction is listed for that enzyme, this is taken as the reaction direction.

When there are multiple enzymes, a single entry in the **S** matrix is required to avoid a combinatorial explosion in the number of elementary modes [Palsson (2006)]. Its direction is assigned as follows.

The reaction is taken as reversible if (i) any of the enzymes is given as reversible, even if others are unidirectional or unknown; or (ii) there are unidirectional enzymes for both directions. If no direction is listed for one of a multiple enzyme set, but there is another enzyme for the same reaction given as unidirectional, the conservative preliminary assumption is made that the unknown enzyme runs in the same direction, subject to revision if this leads to a conflict with pathway information.

The database lists a large number of empirical pathways; all such pathways that contain a particular reaction are now examined and as for enzymes, the "pathway-associated" direction is taken as reversible if it is either reversible in a single pathway or in opposite directions in different pathways.

Next, the enzyme derived directions are reconciled with the "pathway" directions. If the pathway direction for a reaction is unique but the enzyme assignment is reversible, it is taken as reversible because the empirical pathway list is not exhaustive. If there is no direction assigned from the enzymes, the pathway assignation is taken. If neither source indicates a direction, the listed reaction specification is accepted. That would for

example apply to simple, uncatalysed reactions. Finally, if there is a conflict between unidirectional assignments from enzymes and pathways, this is assumed to arise from the preliminary assumption. I.e., one of the alternative enzymes with unknown direction or even an unknown enzyme must catalyse the observed opposite direction and so the reaction is taken as reversible in this case.

These assignment rules are admittedly only an attempt to deal rationally with incomplete or ambiguous information in the database. The underlying rationale is that whenever a unique direction is assigned to a reaction that is taken; but if there is room for doubt, the most liberal assumption, that the reaction is reversible, is made. It is possible that this may lead to spurious modes, and it is advisable that all modes involving reversible reactions should be inspected carefully. But this is in keeping with the general philosophy of constraints-based modelling: the goal is not to uniquely predict a network state, but rather to narrow down feasible states to those that are compatible with known constraints. In practice, the problem did not arise in the flavonoid subnetwork discussed below as despite the liberal interpretation, no reversible reactions were actually found in the subnetwork.

The final type of information required is the classification of compounds as internal or external, for partitioning the **S** matrix. To facilitate this a separately compiled list of acceptable external metabolites is set up. For the complete network, it consists of molecules or groups exchanged with the environment such as H_2O, O_2, CO_2, sulfates, inorganic phosphates, etc. For subnetworks the list is further extended as described in the next section.

For each compound encountered in the reaction listing, all reactions in which it participates are inspected. It is taken as internal if there is a reaction containing it on the left as well a reaction that contains it on the right. If it only occurs on the left it is an external substrate, and an external product if it only occurs on the right. Any compound classified as internal in this way but which also appears on the external metabolite list is then reclassified as a "free" external compound, i.e. it may be either absorbed from or delivered to the environment. Inconsistencies between the compounds identified as external from the network and those on the exchange compounds list generally need to be investigated, as they indicate that either the network specification or the list is incomplete.

3. SUBNETWORKS

Flux balance analysis of the complete known genome-scale network has been undertaken for prokaryotes (single cell organisms) such as *Eschericia Coli* [Reed et al. (2003)] *Heliobacter Pylori* [Schilling et al. (2002)] and yeast [Forster et al. (2003)]. However, even for a relatively small representative network for E. Coli, consisting of 112 reactions and 89 compounds, a very large number of 2.4 million elementary modes are calculated [Gagneur and Klamt (2004)]. By comparison, the complete network for *Arabidopsis* constructed according to the previous section has 1178 reactions and 1089 metabolites which is clearly intractable. Moreover, in a study focussed on a particular aspect of metabolism, one is only interested in modelling the relevant part of the network and would prefer to avoid being distracted by unrelated biological processes.

In order to focus the subnetwork on the area of specific interest (flavonoids), AraCyc is first searched for all the empirically determined pathways that are listed as representing flavonoid production, and the reactions constituting these pathways collected as a preliminary subnetwork. Then, the specification of each of these pathways is inspected for references to additional feeder pathways or reactions, and those are added to the subnetwork. This process is iterated until convergence, and needs only 4 iterations to yield a prototype flavonoid subnetwork of 71 reactions that involve 91 different compounds, of which 44 are internal.

However, to establish if this is a well isolated subnetwork, a number of points need to be investigated. First, compounds that are external to the subnetwork but not external to the complete network, have implied reservoirs outside of the subnetwork and this needs to be justified by identifying known biological or biochemical processes that supply them. For example, a number of the external compounds actually found from the flavonoid prototype, belong to the group of 11 compounds known to form the anabolism-catabolism interface. Others are well known carrier molecules arising e.g. from photosynthesis, such as ATP, ADP (used for energy transfer), NAD and NADH (redox reactions), molecular cofactors, etc. Others, like nucleotides and amino acids are essential for primary metabolism and so can be assumed available for secondary processes such as flavonoid production. A second point is that some such known external compounds may appear as internal from the automated classification since they appear on both sides of subnetwork reactions. However there is no reason to suppose that they

need to be mass balanced within the subnetwork and so have to be explicitly reclassified.

Both of these points are dealt with in this study by adding all compounds for which there is a known reservoir, to the previously mentioned external metabolite list for the complete network.

Even if a fully consistent internal/external classification is achieved in this way, there is no guarantee that the prototype subnetwork contains all reactions in the original network connected to those in the subnetwork.

One way to check that, is to take the full network and cut it at all nodes that have been classified as external in the prototype. Such cuts would be expected to separate the full network into isolated sections. If any of these sections is identical to the prototype, that would confirm its completeness. Possibly it may be distributed over more than one section, in which case such sections together would constitute the complete subnetwork. Otherwise, if a section can be identified that contains the prototype as a subset, any additional reactions and compounds could be added to the subset and/or additional cuts identified that would separate off a smaller section containing the prototype.

This idea has been elaborated as a method to split a network into subnetworks without a priori identification of an area of interest, by Schuster et al. (2002). In that approach, all compounds that are represented by nodes with a connectivity higher than a threshold value, are reclassified as external. The rationale is that if a compound takes part in many reactions, it needs not be conserved along any one path individually and so is "operationally buffered". A threshold value of 4 or 5 is typically used. Most of the previously mentioned external compounds are readily identified in this way, e.g. carrier molecules have connectivities of the order of 100 or more in the AraCyc network. A software implementation of the network splitting algorithm is available under the name SEPARATOR from http://pinguin.biologie.uni-jena.de/bioinformatik/networks/index.html.

Applying this to the metabolism of the bacterium *Mycoplasma pneumoniae*, Schuster et al. (2002) find that the full network is decomposed into 19 subnetworks with identifiable individual biochemical functions. Similarly useful splittings are discussed for the human redox metabolism by Schwarz et al. (2005) .

However for the Arabidopsis network, no suitable decomposition could be obtained by connectivity splitting. Using the externals as identified from the prototype subnetwork, the full network splits into a total of around 200 subnetworks. The vast majority (about 90%) of these consist of only a single reaction; these are mostly already isolated reactions in the original AraCyc, probably reflecting incomplete knowledge of the network. At the other extreme is a single large block, that contains about 75% of all reactions. That leaves about 10% of all reactions distributed over about 20 fragments, most of which contain 5 or less reactions, and a few in the range of 6 – 30 reactions. Parts of the flavonoid prototype network are recognisable in one or two of these midsize fragments, but at least half of it remains buried in the large unresolved block. Experimenting with connectivity threshold values changes the exact numbers but does not change the overall picture.

While this appears to indicate that the flavonoid subnetwork is simply inextricably linked to the rest of the *Arabidopsis* metabolism, a further attempt was made to disentangle it by applying the following heuristic strategy.

- The classification of compounds described in section 2 is applied separately to the full network of balanced equations and the prototype flavonoid network, using the same list of "approved" external metabolites as compiled for the prototype, for both.

- Then all compounds that are external in the subnetwork but internal in the full network, are identified. For each of these, a pathway in the full network that will connect it only to known external metabolites is traced back manually. All reactions encountered along the way are added to the subnetwork. This may result in adding new internal compounds as well, also processed iteratively.

- Finally the connectivity of each compound that is internal in both networks is calculated, giving values c_f and c_s for the full and subnetwork respectively. By definition $c_f \geq c_s$. If they are equal, all relevant reactions from the full network are already included in the subnetwork. If $c_f >> c_s$, that usually means that the compound can be added to the externals list because there are enough processes outside the subnetwork to account for its conservation. If $c_f - c_s$ is small, similar tracebacks as in the previous point are performed and by adding all reactions encountered, it is ensured that $c_f = c_s$ for all internal compounds.

If this strategy succeeds, the resulting extended subnetwork fulfils the criteria that a) all internal nodes are disconnected from the remainder of the full network, and b) all boundary nodes represent compounds for which either the existence of a reservoir is justifiable on biochemical grounds, or if not, its status is external in the full network as well. This means that the inconsistency does not arise from the construction of the subnet and needs to be resolved by additional information on the complete metabolic network.

However, the strategy is clearly only a heuristic and there is no guarantee that it will succeed. In particular, it may well happen that the traceback step diverges in a growing cascade of reactions being added to the subnet that only terminates when all or a large part of the full network has been added to it.

4. RESULTS AND DISCUSSION

The outlined strategy to augment the prototype is surprisingly effective when applied to the extraction of a flavonoid subnetwork from AraCyc 3.0. A subnetwork that is fully coherent according to the criteria stated above is obtained with only a 25% increase in size from the prototype (71 reactions increasing to 89, and 91 metabolites to 115). The resulting network structure is shown in Figure 1 and is readily amenable to elementary mode analysis, as described in a subsequent paper Clark and Verwoerd (2007). Similar success was achieved there with AraCyc 3.5, yielding a somewhat larger and more complicated subnet.

As an independent check, the SEPARATOR program mentioned in section 3 was again run on the full network, but specifying external metabolites according to the outcome of the augmented subnetwork. As before, a large block of 824 reactions and 162 single reactions accounts for most of the network. Between these extremes, there is one block each of 86, 12, and 10 reactions, while the remaining 27 subnets have sizes between 2 and 6. Significantly, the 86 reaction block together with two small fragments are identical with the heuristically constructed subnet. This confirms that no further reactions are needed to complete this subnet.

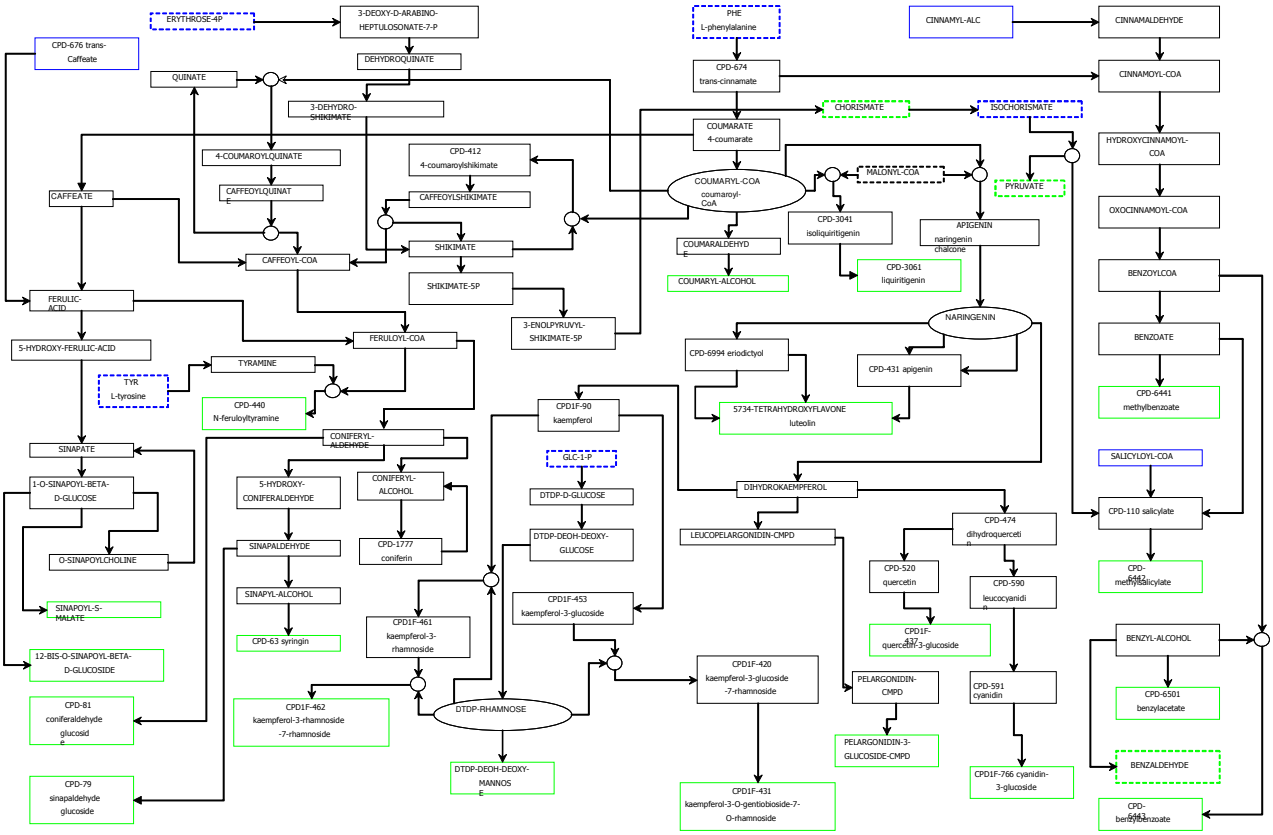

Figure 1: Flavonoid metabolic reaction subnetwork for Arabidopsis Thaliana extracted from AraCyc V3.0. External reactants are shown in blue boxes, and external products in green. Dotted outlines indicate compounds identified by the external compound listing. For simplicity, only the backbone network is shown; most reactions also involve secondary (exchange) metabolites not shown in the diagram.

In contrast the connectivity threshold method was able to partition the simpler metabolic networks mentioned before, more uniformly into a range of functional units. Although that method did not work for *Arabidopsis*, it is not clear whether a more sophisticated method might still yield a single set of externals that will partition the Arabidopsis network similarly. It seems likely however that this network is simply too complicated for that to be possible.

5. CONCLUSION

A general method to decompose a metabolic network should ideally produce a set of roughly similarly sized subnets with identifiable distinct biochemical functions. It has been demonstrated elsewhere that connectivity threshold cutting can achieve that for some metabolic networks.

The method outlined in this work has a more modest goal, to isolate just a single subnet based on ad hoc identification of a core phenomenological pathway connected to some specific biological function. It is difficult to predict whether even this is possible in a specific case, but application of the heuristic proposed here did achieve that goal for flavonoid production in *Arabidopsis* whereas the threshold method did not yield a useful outcome. Further experience is needed to test its general applicability.

6. REFERENCES

Atkinson, D. E. (1977), Cellular energy metabolism and its regulation, Academic Press New York.

Clark, S. and W. S. Verwoerd (2007). Using a reconstructed flavonoid subnetwork to study anthocyanin biosynthesis. MODSIM07, Christchurch.

Enright, A. J., S. Van Dongen and C. A. Ouzounis (2002), An efficient algorithm for large-scale detection of protein families, *Nucl. Acids Res.* 30(7), 1575-1584.

Forster, J., I. Famili, P. C. Fu, B. O. Palsson and J. Nielsen (2003), Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network, *Genome Research* 13, 244-253.

Gagneur, J. and S. Klamt (2004), Computation of elementary modes: a unifying framework and the new binary approach, *BMC Bioinformatics* 5, 175.

Karp, P. D., C. A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahren, S. Tsoka, N. Darzentas, V. Kunin and N. Lopez-Bigas (2005), Expansion of the BioCyc collection of pathway/genome databases to 160 genomes, *Nucleic Acids Research* 19, 6083-6089.

Klamt, S., J. Saez-Rodriguez and E. Gilles (2007), Structural and functional analysis of cellular networks with CellNetAnalyzer, *BMC Systems Biology* 1(1), 2.

Palsson, B. O. (2006), Systems Biology - Properties of Reconstructed Networks, Cambridge University Press, 322 pp., New York.

Reed, J. L., T. D. Vo, C. H. Schilling and B. O. Palsson (2003), An expanded genome-scale model of Eschericia coli K-12 (iJR904 GSM/GPR), *Genome Biology* 4, R54.1-R54.12.

Rhee, S. Y., W. Beavis, T. Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L. A. Mueller, S. Mundodi, L. Reiser, J. Tacklind, D. C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon and P. Zhang (2003), The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community., *Nucleic Acids Research* 31(1), 224-228.

Schilling, C. H., M. W. Covert, I. Famili, G. M. Church, J. S. Edwards and B. O. Palsson (2002), Genome-scale metabolic models of less-characterised organisms: A case study for Heliobacter pylori, *Journal of Bacteriology* 184, 4582-4593.

Schuster, S., T. Pfeiffer, F. Moldenhauer, I. Koch and T. Dandekar (2002), Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae *Bioinformatics* 18(2), 351-361.

Schwarz, R., P. Musch, A. von Kamp, B. Engels, H. Schirmer, S. Schuster and T. Dandekar (2005), YANA - a software tool for analyzing flux modes, gene-expression and enzyme activities, *BMC Bioinformatics* 6(1), 135.